ABSTRACT OF THE DISCLOSURE

An improved duplicate detection technique that uses query-relevant information to limit the portion(s) of documents to be compared for similarity is described. Before comparing two documents for similarity, the content of these documents may be condensed based on the query. In one embodiment, query-relevant information or text (also referred to as "snippets") is extracted from the documents and only the extracted snippets, rather than the entire documents, are compared for purposes of determining similarity.